Recognizing Faces and Human Gestures in Natural Expressions by Considering Computer Vision Techniques

M Ashok Kumar[1], Dr.R.R.Tewari[2],
[1]Department of Information Technology,V R Siddhartha Engineering College,Kanuru, Vijayawada, India,
Ashokkumar.munnangi@gmail.com
[2]Professor in Department of Computer Science, J.K.Institute of Applied Physics and Technology,
University of Allahabad-211002

Abstract— PCs ought to have the capacity to perceive feelings, by dissecting the human's emotional state, physiology and conduct. Keeping in mind the end goal to make human-PC interfaces genuinely characteristic, we have to create innovation that tracks human development, body conduct and outward appearance, and translates these developments in a viable way. Individuals normally convey what needs be through facial motions and expressions. We will probably fabricate a facial motion human-PC interface for use in robot applications. Our framework does not require exceptional brightening or facial cosmetics. By utilizing numerous Kalman channels we precisely anticipate and powerfully track facial components. This is in spite of unsettling influences and fast developments of the head (counting both translational and rotational movement). Since we dependably track the face continuously we are likewise ready to perceive movement motions of the face. Our framework can perceive a huge arrangement of motions (13) running from "yes", "no" and "may be" to distinguishing winks, flickers and resting.

Keywords— facial expression, gesture, recognition, multimodal interface

I. INTRODUCTION

Applications including human robot interfaces with cutting edge association abilities have begun to get impressive consideration in the scholarly group in modern research centers and in the media. A percentage of the best experimental difficulties towards such applications are identified with the advancement of suitable innovations and methods for robots to see people and track their action. Following of hands development gives data to hand-motion acknowledgment frameworks, while face and facial elements encode basic data about outward appearance and head development.

Hands and confronts assume an imperative part for human correspondence. They are the fundamental wellspring of data to separate and recognize individuals, to translate informative signs as hand and face signals and to comprehend feelings and goals in light of outward appearances.

The condition of every item is thought to be an imperceptibly Markov process which advances as per particular progress and which creates estimation expectations that can be assessed by contrasting them and the real picture estimations. Model-based methodologies are computationally more costly and frequently require the appropriation of extra imperatives for the progress of the framework and for the Believe ability of every posture except they intrinsically give wealthier data in regards to the real stance of the followed human and also the correspondence of particular body parts with the watched picture.

A standout amongst the most vital commitment of this paper is identified with the advancement of an incremental classifier which expands the above-portrayed blob following methodology and which is utilized to keep up and ceaselessly redesign a conviction about whether a followed speculation relates to a facial area, a left hand or a right hand. For this reason, we utilize a straightforward, yet vigorous list of capabilities which passes on data about the state of each followed blob, its movement qualities and its relative area as for different blobs.

The reason for the above-depicted methodology for hand, face and facial elements following is to bolster acknowledgment of hand motions and outward appearances for rich communication with a self-governing portable robot. It has been incorporated into a framework which keeps running progressively on a routine PC which is situated on the versatile robot itself.

In this paper, we introduce an examination of hand-over-face motions in a naturalistic video corpus of complex mental states. We characterize three hand-over-face signal descriptors, in particular hand shape, hand activity and facial area impeded and propose a philosophy for programmed discovery of face impediments in recordings of regular appearances.

2. The Vision System

We utilize the MEP following vision framework to execute our facial signal interface. This vision framework is produced by Fujitsu and is intended to track in genuine time different layouts in the edges of a NTSC video stream. It comprises of two VME-transport cards, a video module what's more, a following module which can track up to 100 layouts all the while at video casing rate (30Hz for NTSC). A MC68040 processor card running Vx Works executes

the application program and controls the vision framework.

To track a layout of an article it is important to figure the bending not just at one point in the picture however, at various focuses inside of the inquiry window. To track the development of an item the following module discovers the position in the picture outline where the layout matches with the most reduced bending. The movement is spoken to by a vector to the starting point of the most reduced mutilation.

By moving the inquiry window along the hub of the movement vector items can be effectively followed. The following module performs up to 256 cross relationships for every format inside of a pursuit window.

Issues emerge when the vision framework is utilized to track a face in a head and shoulders picture of a man. Since the head involves the majority of the picture, one format of the whole face surpasses the most extreme layout size reasonable in the vision framework. In this manner, it is just conceivable to track individual elements of the face, for example, the eyes or mouth. Through experimentation we have found that facial highlights with high complexity are great applicants as following formats. For instance an eyebrow which has all the earmarks of being a dim stripe on a light foundation (if the individual has light skin) and the iris of the eye which shows up as dim spot encompassed by the white of the eye are appropriate for following. On the other hand, some facial elements are not as simple to track. For instance the edges of the mouth are hard to track. This is on the grounds that these components are made up fundamentally of plain skin (80%) and the relationship with a facial layout of just plain skin is not altogether distinctive i.e. yields a low twisting.

These issues are further convoluted by the way thatappropriate following components can change their appearance significantly when a man moves their head. The shad-ing of the components can change because of uneven illumi-country and the elements seem to distort when the head is turned, climbed, down or tilted to the side. Every one of these progressions build the twisting regardless of the fact that a layout is coordinating decisively at the right position. It likewise brings about low twists at the wrong arranges which then cause the hunt window to be erroneously moved far from the component. This issue emerges when a head is turned adequately sufficiently far for one portion of the face with all its related components to totally vanish. When the following element has left the pursuit window the development vectors ascertained by the vision framework are eccentric. We have built up a technique to permit a pursuit window to effectively locate its lost component in this manner yielding a dependable face tracker.

Figure 1: Facial Tracking Features

3. Coding Of Hand-Over-Face Gestures

Serving as an initial phase in programmed characterization, we coded hand-over-face motions utilizing an arrangement of descriptors. In this segment, we depict the decision of the dataset, the coding composition, the naming, annotation evaluation and how we create the ground truth names that are utilized as a part of our machine learning tests.

3.1 Dataset

The main test was to discover a corpus of recordings of natural expressions. Since the vast majority of the work on influence investigation concentrates on the face, the majority of the openly accessible common datasets likewise concentrate on countenances with restricted or no impediment. Since we are occupied with the transient part of the hand motion too, still photo corpora did not fulfill our criteria. The freely accessible Cam3D [20] has regular appearances and does not limit the video gathering to confronts. It incorporates abdominal area recordings that have hand-over-face impediments in around 25% of the recordings. The expressions in Cam3D are inspired as a major aspect of a feeling elicitation test, which infers that the hand motions communicated are well on the way to be a piece of articulation of feeling. We are keen on distinguishing such possibly educational signals.

In Cam3D, division is occasion based, so every video section contains a solitary activity. The dataset has 192 video sections that contain hand-over face impediments. These recordings originate from 9 members with mean term of the video being 6 seconds. We utilized the majority of the blocked recordings. For equalization, we likewise arbitrarily chose another 173 video fragments from the Cam3D dataset that don't contain face impediments. The picked no-impediment recordings were chosen containing the same 9 members while keeping the quantity of tests per every member as adjusted as could be expected under the circumstances. This prompted an arrangement of 365 recordings altogether.

3.2 Labelling

With a specific end goal to continue to programmed discovery, we expected to code the hand-over-face impediments present in the dataset. The objective was to code hand motions as far as sure descriptors that can depict the motion. Propelled by the coding construction gave by Mahmoud et al. [22], we coded the signals as far as hand shape, hand activity and facial locale impeded.

Marking was completed utilizing Elan video annotation apparatus [19]. Two master coders (analysts in our exploration gathering) were told to mark the recordings given the accompanying directions:

•*Hand Action:* coded as one label for the whole video according to the action observed in the majority of the frames. Labels are: 1) Touching - If the hand isstatic while touching the face. 2) Stroking/tapping- repetitive motion of the hand on the face. 3) Sliding- any other hand motion that is not repetitive.

• *Hand Shape:* coded as one mark for each casing. It portrays the state of the hand on the face in a particular casing. Marks are fundamentally unrelated, i.e. one mark is allowed per outline. Marks are: 1) Fingers or any different fingers. 2 Open Hand(s) or palm(s). 3) Closed Hand(s) or a clench hand shape. 4) Hands Together - tangled hands.

• *Facial Region Occluded:*coded as one - or numerous - names per edge (names are not commonly exclusive).It depicts the face zone secured - or halfway secured - by the hand amid impediment. Names are: 1)Forehead. 2)Eye(s). 3)Nose. 4)Cheek(s). 5)Lips. 6)Chin. 7)Hair/ear.

3.3 Coding assessment & refinement

To evaluate the coding blueprint and pick up trust in the marks acquired, we ascertained between rater assention between the two master annotators utilizing time-cut Krippendorff 's alpha [16], which is broadly utilized for this sort of coding appraisal in light of its autonomy of the quantity of assessors and its vigor against defective information [14]. When we investigated the reason of the contradiction in these classes, this was for the most part due to the not very many specimens accessible of these classifications in the dataset, for instance: eyes, brow and hair/ear areas had just 25, 100 and 10 edge tests separately, i.e. under 0.2% of the aggregate number of edges altogether.

We chose to prohibit these classifications (for the most part upper face range) in the machine learning venture, as it was out of line to attempt to naturally learn and characterize these classes when the human annotators neglected to concur. Because of the way of our uneven dataset, a few names had not very many specimens. In the arrangement stage, we chose to total a percentage of the gatherings together. The nose area was consolidated with the cheek district as one descriptor of the center face locale. For the hand activity descriptor, we joined sliding, stroking and tapping in one gathering speaking to non-static hand motion, i.e. any sort of movement .

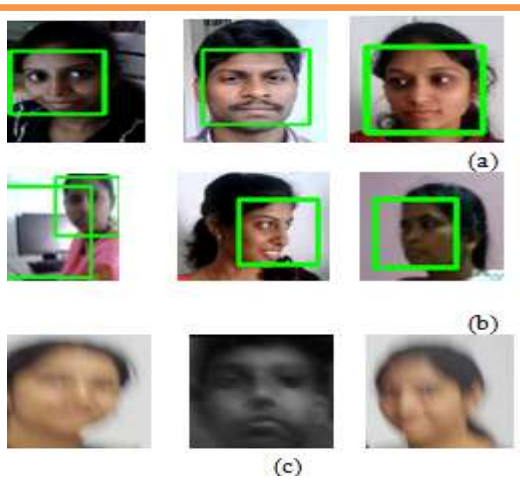


Figure 2: Sample frames from videos in the datasetCam3D showing examples of face touches present in the dataset [20]. Note the challenging - close to natural - recording settings like inconsistent lighting conditions and strong head rotations

4. Detection and tracking of facial features

For following individual facial elements inside of each identified facial blob, we use a half and half approach by coordinating an appearance-based indicator and a component based tracker for the eyes, the nose and mouth. The consolidated methodology acquires focal points from both methodologies allowing hearty recognizable proof of the facial elements, right support of highlight IDs among casings, and in addition constant calculations. The outline of the actualized methodology is delineated in Fig. 3 and depends on three stages: (a) starting identification of facial components utilizing an appearance-based indicator, (b) disposal of false positive location by means of the utilization of anthropometric imperatives, and, (c) constant following of the recognized and separated facial elements utilizing an element based technique. An imperative element which influences both the unwavering quality of discovery and the following exactness of facial components is the extent of the identified face blob. As indicated by Tian [29], facial components turn out to be difficult to identify when the face district is littler than athreshold of approximately 70 × 90 pixels. Therefore, the procedure of facial feature detection and tracking is only activated when the face blob satisfies the above size requirements.

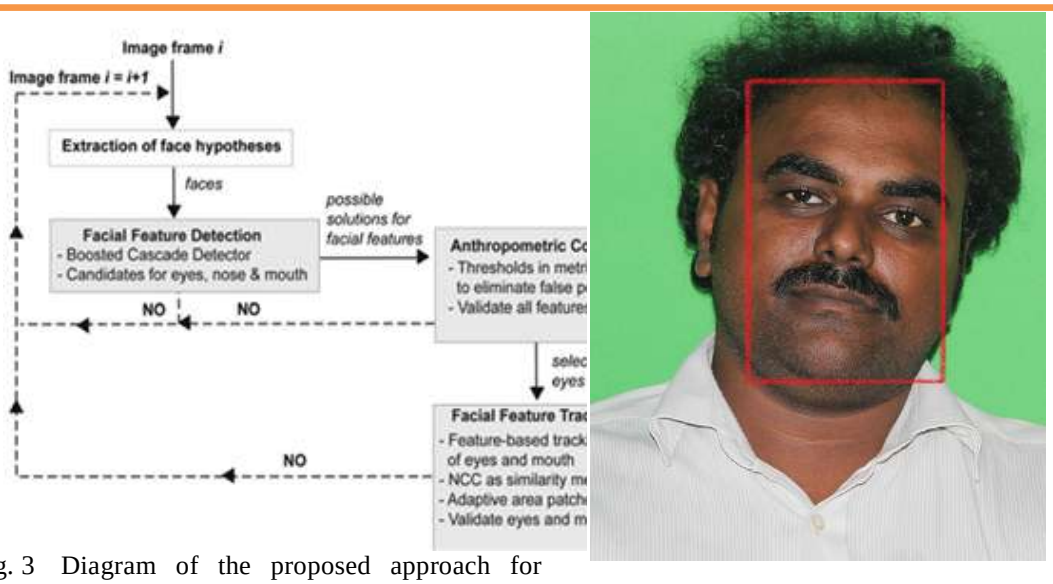


Fig. 3 Diagram of the proposed approach for detection and tracking of facial features

4.1 Detect a Face

First, you must detect the face. Use the vision.CascadeObjectDetector System object™ to detect the location of a face in a video frame. The cascade object detector uses the Viola-Jones detection algorithm and a trained classification model for detection. By default, the detector is configured to detect faces, but it can be used to detect other types of objects.

```
% Create a cascade detector object.
faceDetector = vision.CascadeObjectDetector();

% Read a video frame and run the face detector.
videoFileReader = vision.VideoFileReader('tilted_face.avi');
videoFrame = step(videoFileReader);
bbox = step(faceDetector, videoFrame);

% Draw the returned bounding box around the detected face.
videoFrame = insertShape(videoFrame, 'Rectangle', bbox);
figure; imshow(videoFrame); title('Detected face');

% Convert the first box into a list of 4 points
% This is needed to be able to visualize the rotation of the object.
bboxPoints = bbox2points(bbox(1, :));
```

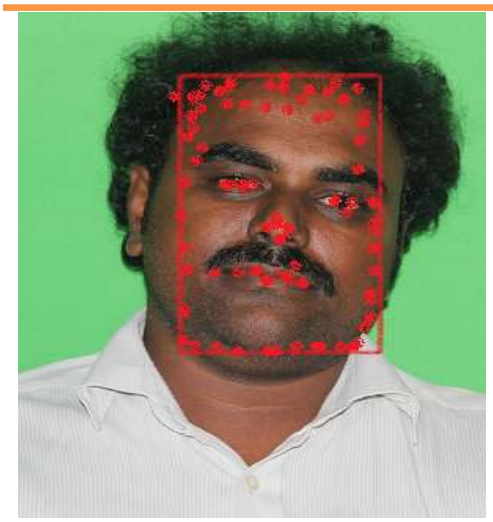

To track the face over time, this example uses the Kanade-Lucas-Tomasi (KLT) algorithm. While it is possible to use the cascade object detector on every frame, it is computationally expensive. It may also fail to detect the face, when the subject turns or tilts his head. This limitation comes from the type of trained classification model used for detection. The example detects the face only once, and then the KLT algorithm tracks the face across the video frames.

4.2 Identify Facial Features To Track

The KLT algorithm tracks a set of feature points across the video frames. Once the detection locates the face, the next step in the example identifies feature points that can be reliably tracked. This example uses the standard, "good features to track" proposed by Shi and Tomasi.

```
% Detect feature points in the face region.
points = detectMinEigenFeatures(rgb2gray(videoFrame), 'ROI', bbox);

% Display the detected points.
figure, imshow(videoFrame), hold on, title('Detected features');
plot(points);
```



4.3 Initialize a Tracker to Track the Points

With the feature points identified, you can now use the vision. Point Tracker System object to track them. For each point in the previous frame, the point tracker attempts to find the corresponding point in the current frame. Then the estimate Geometric Transform function is used to estimate the translation, rotation, and scale between the old points and the new points. This transformation is applied to the bounding box around the face.

```
% Create a point tracker and enable the bidirectional
error constraint to
% make it more robust in the presence of noise and
clutter.
pointTracker =
vision.PointTracker('MaxBidirectionalError', 2);

% Initialize the tracker with the initial point
locations and the initial
% video frame.
points = points.Location;
initialize(pointTracker, points, videoFrame);
```

4.4 Track the Face

Track the points from frame to frame, and use estimateGeometricTransform function to estimate the motion of the face.

```
% Make a copy of the points to be used for
computing the geometric
% transformation between the points in the previous
and the current frames
oldPoints = points;

while ~isDone(videoFileReader)
   % get the next frame
   videoFrame = step(videoFileReader);

   % Track the points. Note that some points may be
lost.
   [points, isFound] = step(pointTracker,
videoFrame);
   visiblePoints = points(isFound, :);
   oldInliers = oldPoints(isFound, :);

   if size(visiblePoints, 1) >= 2 % need at least 2
points

      % Estimate the geometric transformation
between the old points
      % and the new points and eliminate outliers
      [xform, oldInliers, visiblePoints] =
estimateGeometricTransform(...
         oldInliers, visiblePoints, 'similarity',
'MaxDistance', 4);

      % Apply the transformation to the bounding
box points
      bboxPoints = transformPointsForward(xform,
bboxPoints);

      % Insert a bounding box around the object
being tracked
      bboxPolygon = reshape(bboxPoints', 1, []);
      videoFrame = insertShape(videoFrame,
'Polygon', bboxPolygon, ...
         'LineWidth', 2);

      % Display tracked points
      videoFrame = insertMarker(videoFrame,
visiblePoints, '+',  'Color', 'white');

      % Reset the points
      oldPoints = visiblePoints;
      setPoints(pointTracker, oldPoints);
   end

   % Display the annotated video frame using the
video player object
   step(videoPlayer, videoFrame);
end

% Clean up
release(videoFileReader);
release(videoPlayer);
release(pointTracker);
```

More specifically, we define the following criteria:

- All four selected features (eyes, nose, mouth) should be detected.

- The normalized sizes of the two eyes and mouth should be within certain bounds.

- The normalized distance between the midpoint of the eye centers and nose tip should be approximately 0.6. That is D2 /D1≃ 0.6, where D2 is the distance between points P3 and P4 (see Fig. 4).

- The normalized distance between the midpoint of eye centers and mouth center should be approximately≃ 1.2. That is D3/D1≃ 1.2, where D3 is the distance between points (P3 and P5).

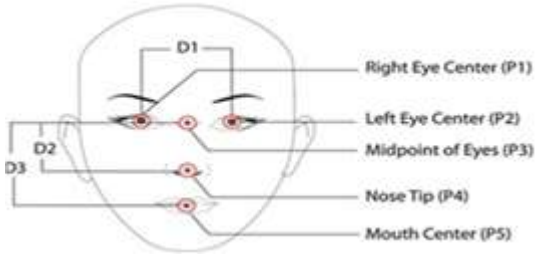


Fig. 4 Landmarks in the anthropometric face model

Tracking is based on template matching, using as eye and mouth templates the detected areas from each face. The ormalized cross-correlation (NCC) measure of Eq. 8 is used as matching score/quality measure.

$$NCC = \frac{(gt - \overline{gt}) \bullet (gs - \overline{gs})}{\| gt - \overline{gt} \| \cdot \| gs - \overline{gs} \|}$$

5. Discussion

Our characterization results for hand location and arrangement got for the six grouping undertakings. The outcomes show generally parallel classifiers with the exception of hand shape where we utilized a 4 class classifier,henceforth the lower grouping qualities. Our multi-modular combination methodology demonstrated a factually critical change over an innocent classifier for the majority of our order experiments. For the testing nature and oddity of the motion characterization undertaking, we consider these outcomes agreeable, considering the way of the uneven dataset we are managing (few preparing tests for a few classifications). Unequal conveyance of the descriptors' qualities among diverse members exhibited a test in the arrangement also. Since our examinations are client autonomous, uneven appropriation of prompts exhibited a test to the classifiers.

6. Conclusion

In this paper, we have introduced a coordinated methodology for following of hands, countenances and facial components in picture successions, proposed to bolster common association with self-governing exploring robots out in the open spaces and, all the more particularly, to give information to the investigation of hand signals and outward appearances that people use while occupied with different conversational states with the robot. For hand and face following,

we utilize a blob tracker which is particularly prepared to track skin-shaded areas. The skin incorporating so as to shade tracker was stretched out an incremental probabilistic classifier which was utilized to keep up and persistently overhaul the conviction about the class of each followed blob which can be a left-hand, a right hand or a face. Facial component recognition and following was performed by means of the job of cutting edge appearance-based identification combined with highlight based following, utilizing an arrangement of anthropometric requirements.

Trial results have affirmed the viability of the proposed methodology demonstrating that the individual favorable circumstances of every included segment are kept up, prompting executions that consolidate precision, proficiency and strength. The motivation behind the proposed following way to deal with encourage human-robot association errands yet the procedure exhibited here has attributes that constitutes it suitable for different undertakings too. Other than

utilizing it to give data for the examination of hand signals and outward appearances, we plan to utilize it for more broad action acknowledgment undertakings and assignments identified with robot learning by exhibition.

References

[1]M. Mahmoud and P. Robinson. Interpreting
hand-over-face gestures. In ACII. 2011.
[2]H. Lausberg and H. Sloetjes. Coding gestural behavior with the NEUROGES-ELAN system. Behavior research methods, 2009.
[3]K. Krippendorff. Content analysis: An introduction to its methodology. Sage, 2004.
[4]A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. Communication methods and measures, 2007.
[5]Sigalas, M., Baltzakis, H., Trahanias, P.: Visual tracking of independently moving body and arms. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS '09), St. Louis, MO, USA, October 2009
[6]Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.: Modelbased hand tracking using a hierarchical bayesian filter. IEEE Trans. Pattern Anal. Mach. Intell. 28(9), 1372-1384 (2006)